Proteomics SIG

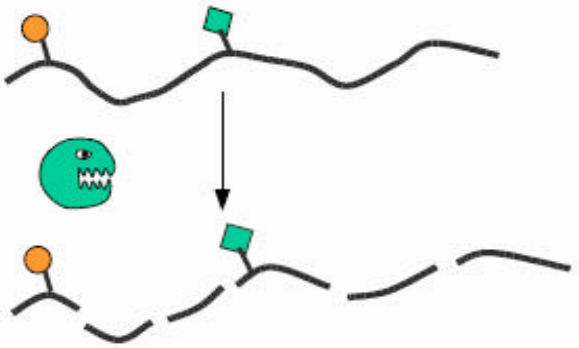
- Spectrometry-based Proteomics
- Standards and Data Exchange

Identification Studies

Spot from 2D gel

Extraction

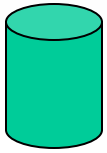
Digestion

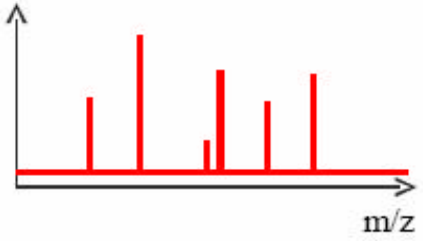


Protein digest database
(MW of all tryptic peptides of all proteins)

MSPQTETK	922.05
ASVGFK	608.72
AGVK	374.46
EYK	439.49
LTYYTPEYETK	1408.55
DTDILAAFR	1022.15
VTPQPGVPPEE	3857.18
YK	310.37
GR	232.26
CYHIEPVPGEE	1996.23

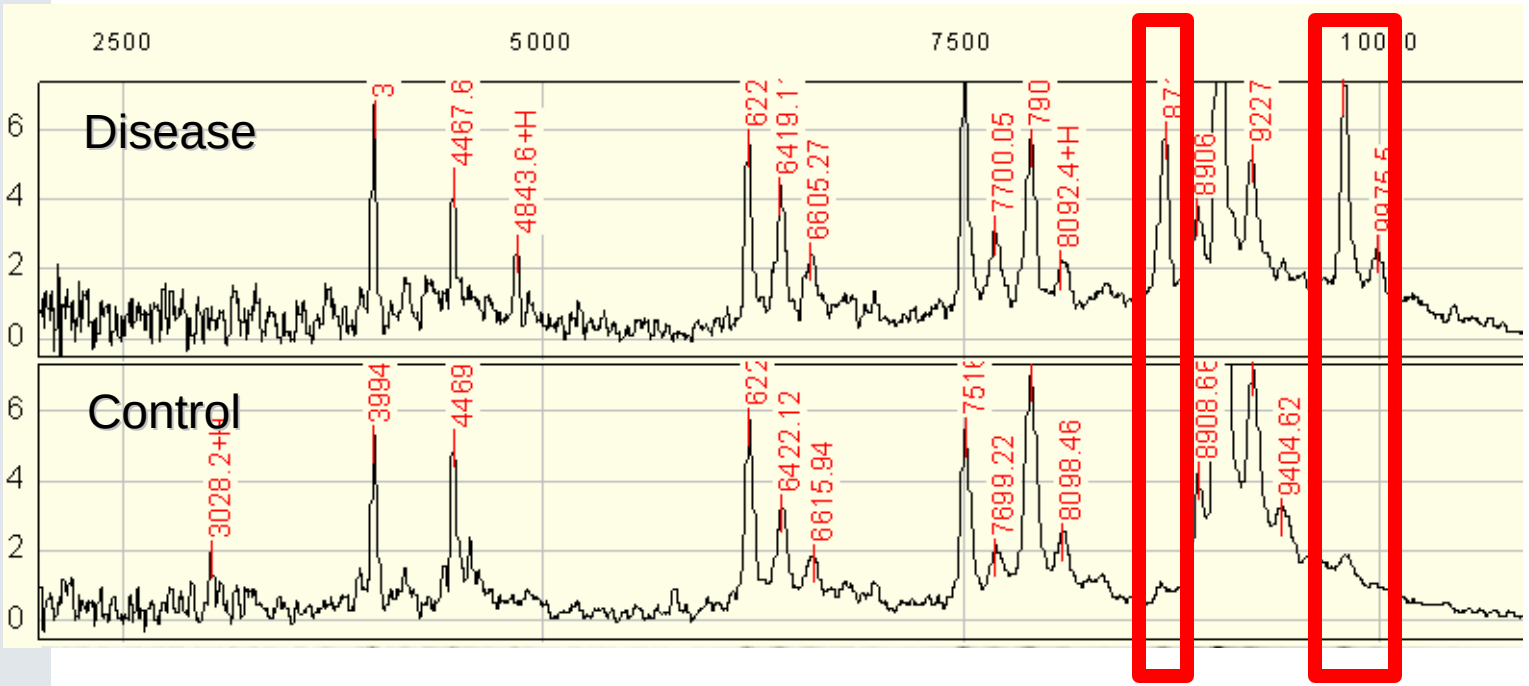
Protein
Database





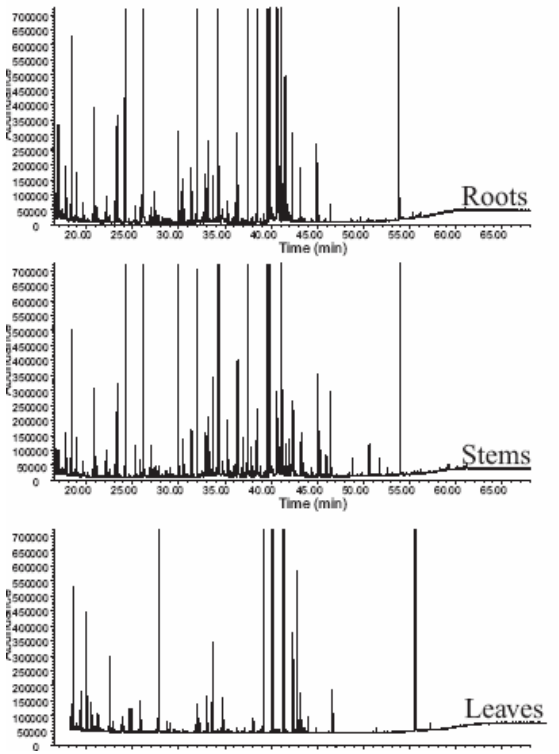
Mass Spec

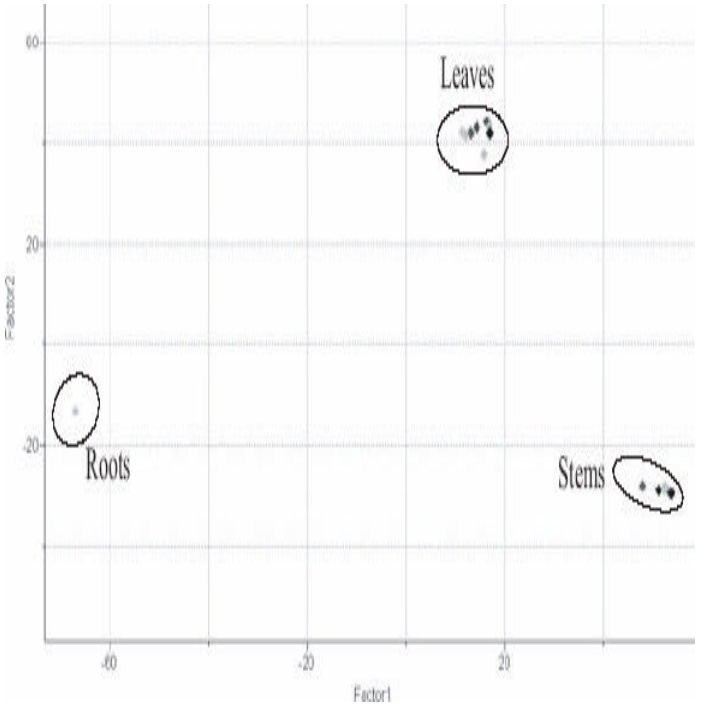
Profiling Studies



Courtesy: Ciphergen

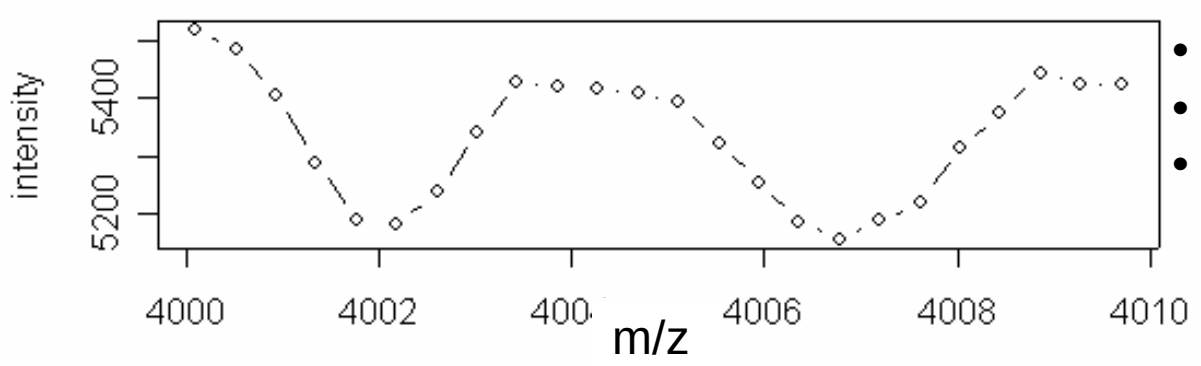
Matabolomics





(Source: Bioinformatics 19: 2283, 2003)

Spectrum Data



- Proteomics
- Metabolomics
- HPLC, IR, etc

❶ Index j	1	2	3	4	...
❷ At j	4000.105	4000.522	4000.939	4001.356	...
❸ Yj	5518.80	5484.58	5406.03	5287.43	...

❶ Index j: also called clock tick, scan #, sample #, variable #

❷ At j: also called m/z, mass

❸ Yj: also called intensity, relative intensity, standardized intensity, abundance

Ontology and Common Data Element

of the optimization methodology, this method can be applied to separations where no prior knowledge of the feed mixture composition exists, which is frequently the case in industrial applications.

Nomenclature

A	user-adjustable weight
a_i	response model parameters
B	user-adjustable weight
COF	chromatographic optimization function
f_i	measure of separation of peak pair i (Figure 2)
g_i	measure of separation of peak pair i (Figure 2)
K_i	peak geometry penalty for peak i
M	number of expected peaks
N	actual number of eluted peaks
n_p	number of peaks exhibited on chromatogram
R_{ij}	resolution between peaks i and j
$\sum R_{ij}$	summation of individual peak pair resolutions over the entire chromatogram
t	total analysis time

tion of Phenylurea Pesticides using Ternary Mobile Phase Gradients in Reversed-Phase HPLC. *J. Liq. Chromatogr.* **1991**, *14*, 3125−3151.

Klein, E.; Rivera, S. Neural Network Signal Interpretation for Optimization of Chromatographic Protein Purifications. *Appl. Math. Computer Sci.* **1998**, *8*, 865−886.

Klein, E.; Rivera, S. A Review of Criteria Functions and Response Surface Methodology for the Optimization of Analytical Scale HPLC Separations. *J. Liq. Chromatogr. Relat. Technol.* **2000**, submitted for publication.

Lindberg, W.; Johansson, E.; Johansson, K. Application of Statistical Optimization Methods to the Separation of Morphine, Codeine, Noscapine and Papaverine in Reversed-Phase Ion-Pair Chromatography. *J. Chromatogr.* **1981**, *211*, 201−212.

Lundell, N.; Markides, K. Optimization Strategy for Reversed-Phase Liquid Chromatography of Peptides. *J. Chromatogr.* **1993**, *639*, 117−127.

Palasota, J.; Leonidou, I.; Palasota, J.; Chang, H.-L.; Deming, S. Sequential Simplex Optimization in a Constrained Simplex Mixture Space in Liquid Chromatography. *Anal. Chim. Acta* **1992**, *270*, 101−106.

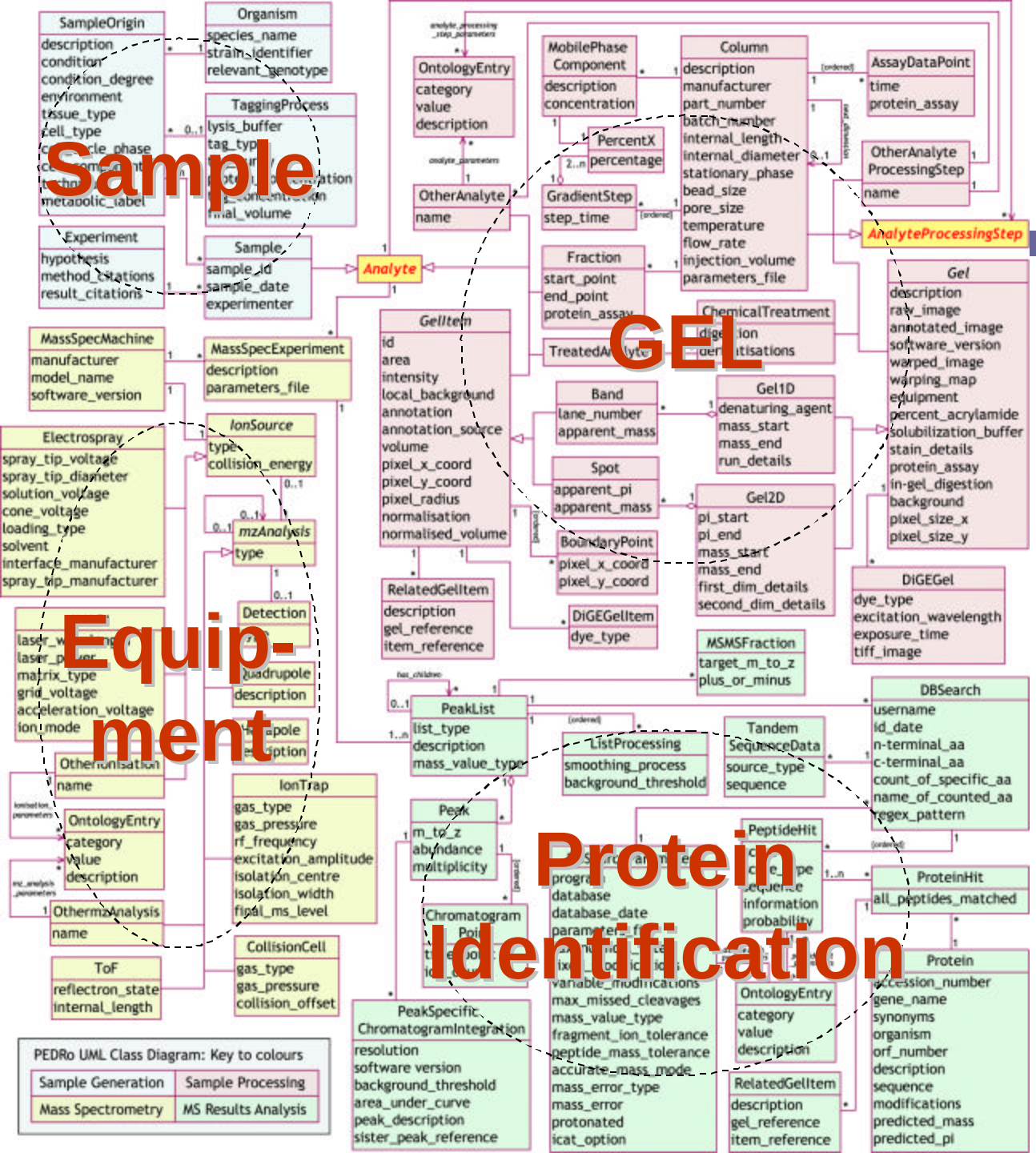
Wang, Q.-S.; Gao, R.-Y.; Yan, B.-W. Computer-Assisted Optimization of pH and Ion Concentration Selectivity in HPLC

Proteomics: Needs and Standards

- Needs
  - Tracking lab work flows -- LIMS
  - Data storage and retrieval -- Proteomic Databases
  - From data to biomarkers – Analytical Algorithms
  - Interoperation among analytical systems – Grid Computing

- Standards
  - MIAPE: Minimum Information About Proteomics Experiments
  - SMOS: Statistical Model Of Spectra (under development at Duke)

MIAPE



caBIG cancer Biomedical Informatics Grid

Statistical Model Of Spectra (SMOS)

- Scope

  Spectrum-based profiling methods, such as proteomics and metabonomic

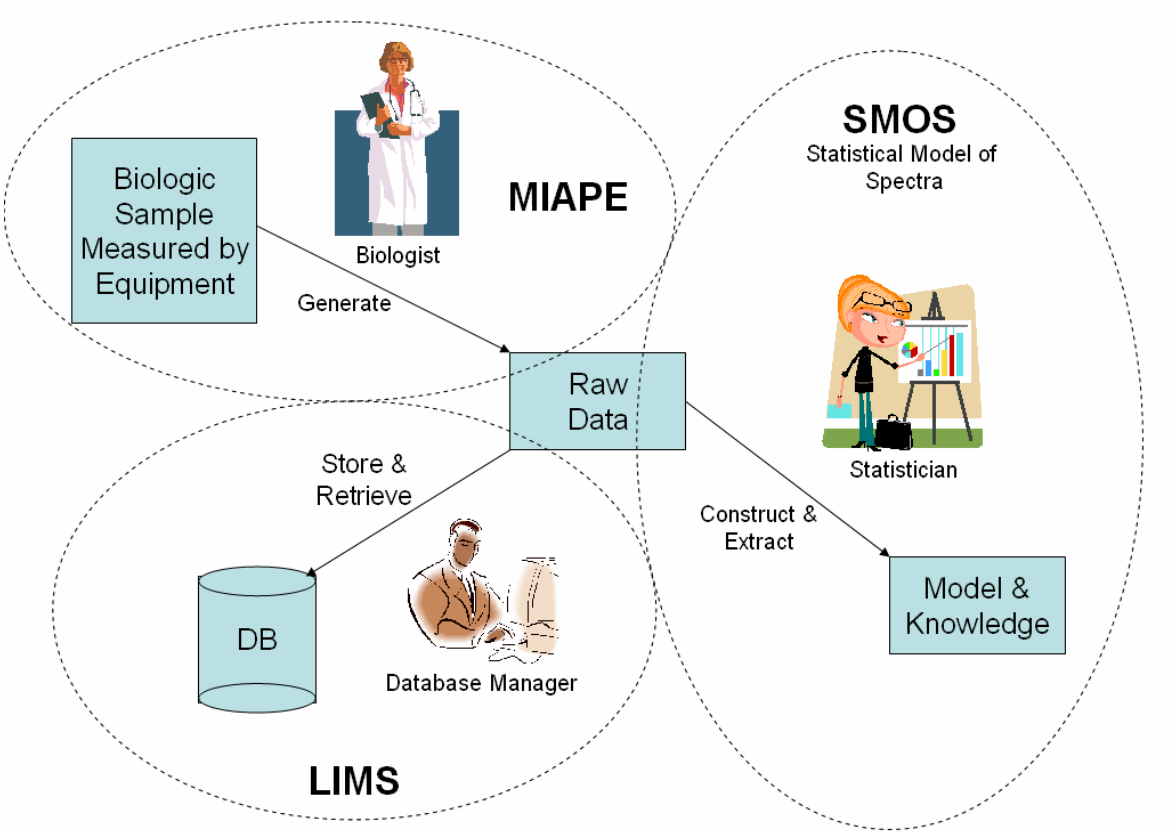
- Focus
  - Statistical modeling
- Purpose
  - Standard for statistical data analysis, exchange, comparison, and verification
  - Audit trail for statistical manipulation of spectral data

Statistical Modeling of Spectra (SMOS)

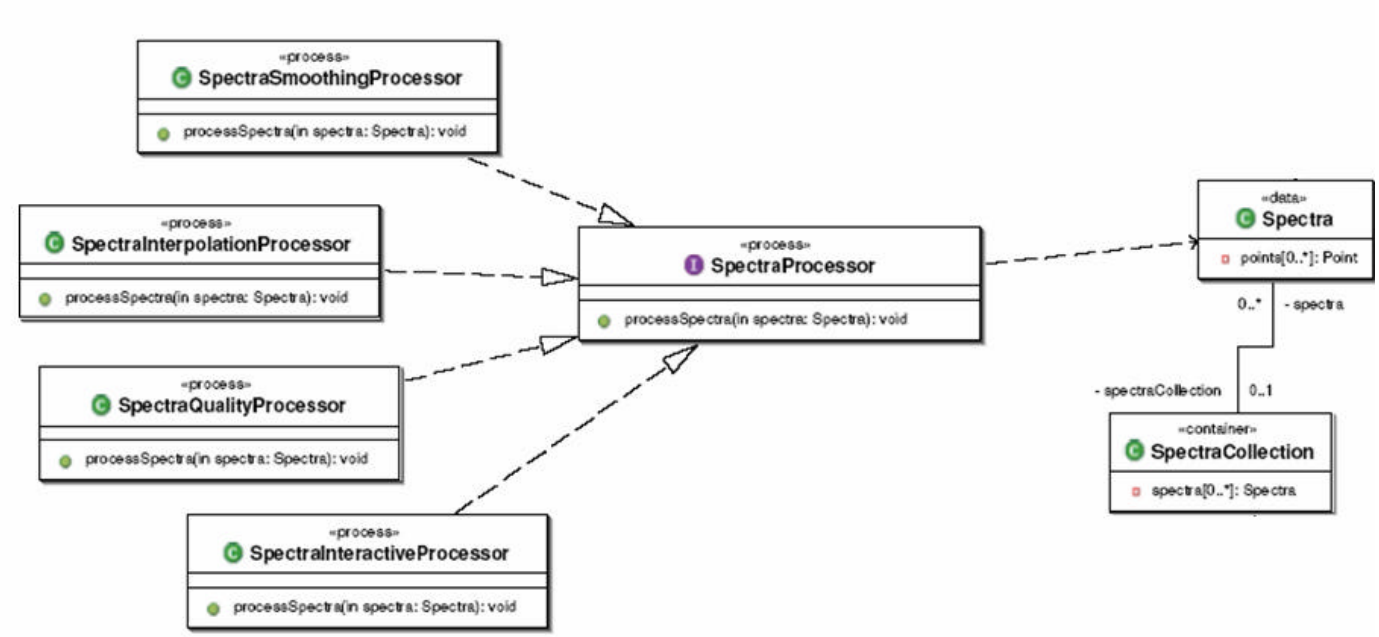
- Single spectrum
  - Baseline removal, Smoothing etc
- A collection of spectrum
  - Normalization, Aggregation, Alignment etc.
- Raw spectrum -> Extracted Features
  - Peaks, Bins, Principle components
- Extracted Features -> Models
  - Clustering, Classification, and Survival
  - Biomarker discovery

[Next: graphic models of SMOS]

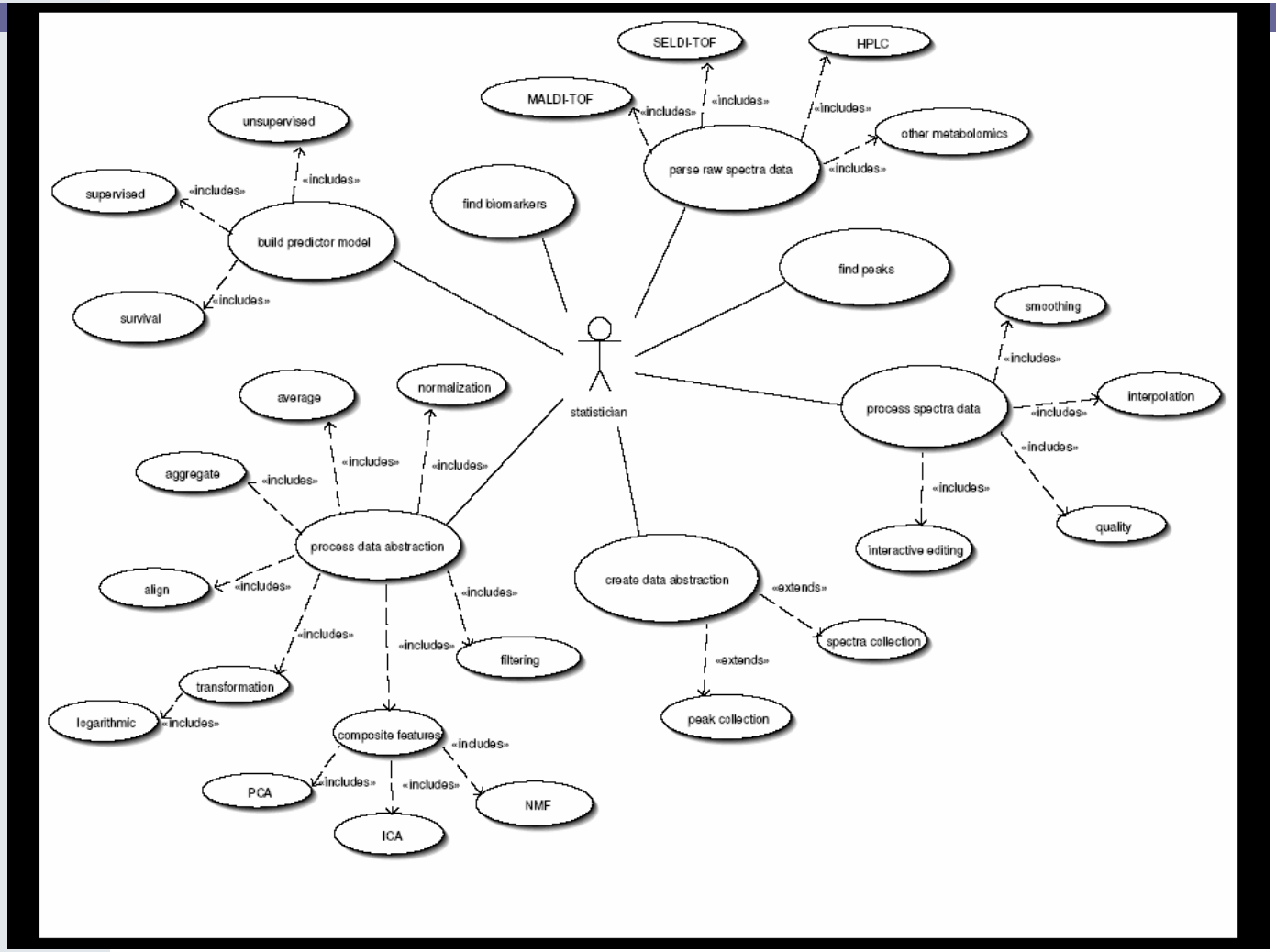
Relationship between MIAPE and SMOS



SMOS: UML model (part)



SMOS: Use case Model



[next: use cases]

Summary of Activities in the Proteomics SIG

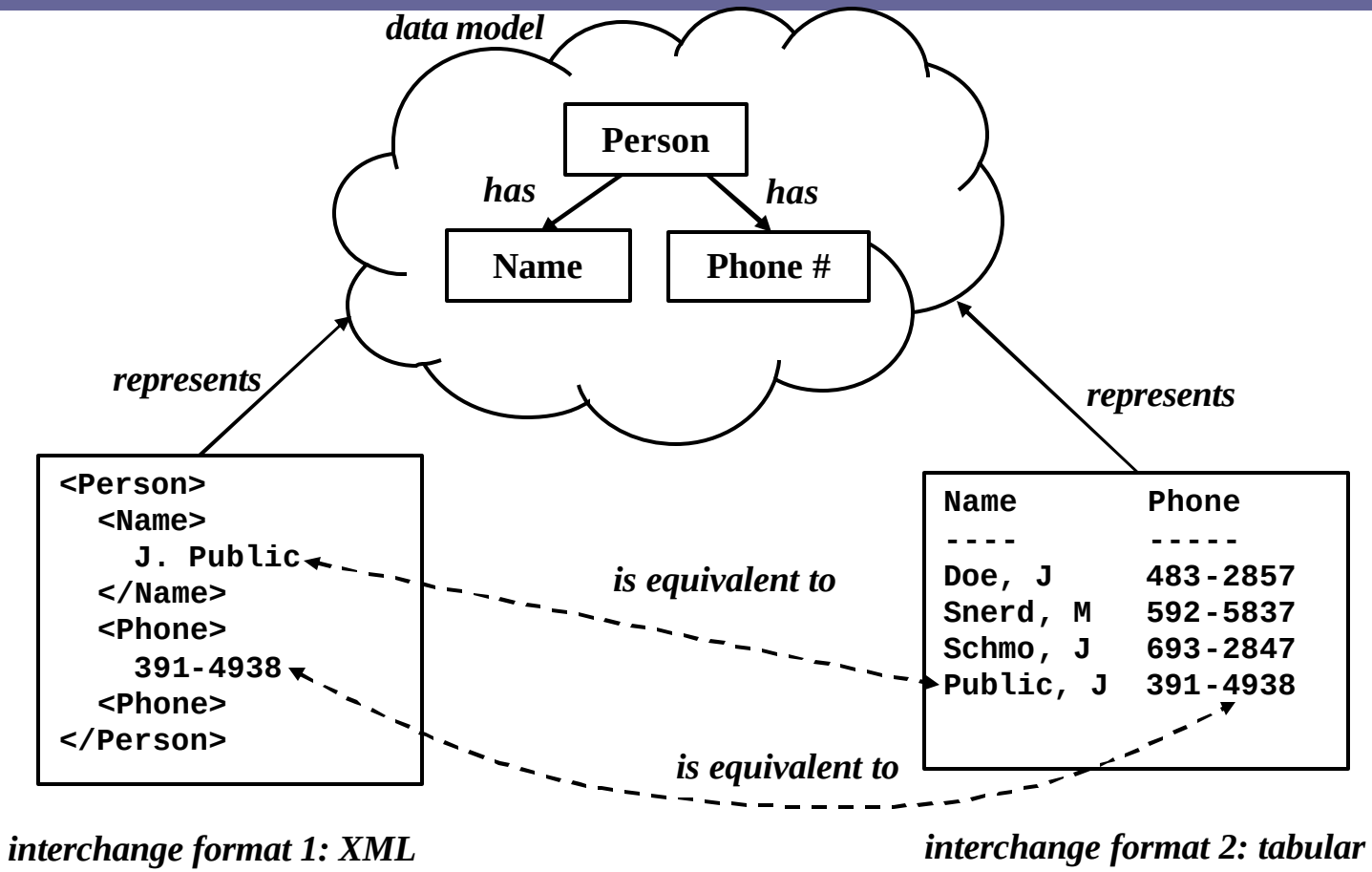
- LIMS – Fox Chase
- Q5 – Dartmouth
- Rproteomics – Duke

Discussion

Data and Metadata

- What is data?
  - Observational (e.g., sensor readings)
  - Computational (e.g., output of simulation)
- What is metadata? – data about data
  - Descriptive (e.g., biological specimen)
  - Relational (e.g., peak table generated from raw spectrum)
  - Contextual (e.g., units of measure)
- What's the difference?
  - Because metadata is still data, the difference could be blurred
  - Metadata is data that helps us use and understand other data

From Data Model to Exchange Format



[slide from NESS]

Data Exchange

- Digital representation of data/metadata model (e.g., file, protocol message)
- Components of an interchange format
  - Syntax
    - Elements (e.g., area code, exchange)
    - Rules governing element types, occurrence, order, cardinality, etc. (e.g., area code is a three-digit integer which precedes the exchange)
  - Representation (e.g., XML, ASN.1, columnar)
  - Encoding (e.g., Unicode, ASCII, binary encoding)